# Phase behavior of disordered proteins underlying low density and high permeability of liquid organelles

## *Supplementary Information*

## Index

### *Supplementary text*



### *Supplementary figures*



## 1. LAF-1 amino acid sequence and RGG sequence properties

The LAF-1 gene was codon optimized for *E. coli*, synthesized by GenScript, and inserted in a pET28a backbone with a N-terminal 6×-His tag. The amino acid sequence of LAF-1 is listed below. The theoretical pI is 6.60. The intrinsically disordered RGG region lies between residues Met1 to Gly168, and is bolded and underline in the sequence below. For a full linear sequence analysis see Fig. S13. Sequence analysis was performed using localCIDER[1]. Here, polar residues are shown in green; positively charged are shown in blue; negatively charged are shown in red; hydrophobic are shown in black; aromatic residues are shown in orange.

```
1    MESNQSNNGG SGNAALNRGG RYVPPHLRGG DGGAAAAASA GGDDRRGGAG
51   GGGYRRGGGN SGGGGGGGYD RGYNDNRDDR DNRGGSGGYG RDRNYEDRGY
101  NGGGGGGGNR GYNNNRGGGG GGYNRQDRGD GGSSNFSRGG YNNRDEGSDN
151  RGSGRSYNND RRDNGGDGQN TRWNNLDAPP SRGTSKWENR GARDERIEQE
201  LFSGQLSGIN FDKYEEIPVE ATGDDVPQPI SLFSDLSLHE WIEENIKTAG
251  YDRPTPVQKY SIPALQGGRD LMSCAQTGSG KTAAFLVPLV NAILQDGPDA
301  VHRSVTSSGG RKKQYPSALV LSPTRELSLQ IFNESRKFAY RTPITSALLY
351  GGRENYKDQI HKLRLGCHIL IATPGRLIDV MDQGLIGMEG CRYLVLDEAD
401  RMLDMGFEPQ IRQIVECNRM PSKEERITAM FSATFPKEIQ LLAQDFLKEN
451  YVFLAVGRVG STSENIMQKI VWVEEDEKRS YLMDLLDATG DSSLTLVFVE
501  TKRGASDLAY YLNRQNYEVV TIHGDLKQFE REKHLDLFRT GTAPILVATA
551  VAARGLDIPN VKHVINYDLP SDVDEYVHRI GRTGRVGNVG LATSFFNDKN
601  RNIARELMDL IVEANQELPD WLEGMSGDMR SGGGYRGRGG RGNGQRFGGR
651  DHRYQGGSGN GGGGNGGGGG FGGGGQRSGG GGGFQSGGGG GRQQQQQQRA
701  QPQQDWWS
```

## 2. Numerical construction of binodals based on Muthukumar's theory for polymers in semi-dilute solutions

We identified a theoretical framework that is able to reproduce the unusual features of measured binodals[2]. This allowed us to uncover the physical underpinnings which give rise to the measured binodals for the RGG domain and for LAF-1 in the presence and absence of RNA. Our numerical approach rests on fitting binodals computed using Muthukumar's theory for semi-dilute polymer solutions[2] to the measured binodals. This numerical analysis takes the measured $B_2$ values as inputs, converts these to Flory $\chi$-values, and extracts free parameters in Muthukumar's theory that include the values for $w$ and the $\chi$-dependent values of $\xi$, such that the parameter estimates yield theoretical binodals that match the experimentally measured ones. Our use of Muthukumar's theory was mandated by our failure to reproduce even qualitative trends of measured binodals when we calculated binodals using the simple mean-field Flory-Huggins expression for describing the thermodynamics of polymer solutions[3] (see Fig. S6). The inclusion of a three-body interaction term with the conventional Flory-Huggins free energy of mixing[4] yields theoretical binodals with narrower two-phase regimes when compared to the form of the theory that ignores three-body interactions. Although this behavior trends in the right direction, there are several features of the measured binodals that cannot be accounted for without further generalizations. These include, the surprisingly low volume fractions associated with high concentration arms of measured binodals, the invariance of the critical point of LAF-1 to the addition of RNA, near coincidence of the low and high concentration arms of binodals for LAF-1 and RGG in the absence of RNA with a lowering of the critical point for RGG, and the independent movement of the high concentration arms of LAF-1 binodals that were measured upon the addition of RNA. These features cannot be reproduced in simple mean field theories that ignore conformational and chain density fluctuations, which are the characteristic hallmarks of polymers in semi-dilute solutions. The full expression for the free energy of mixing in Muthukumar's theory[2] is as follows:

$$\frac{\Delta G_m}{k_B T} = \frac{\phi}{r}\ln(\phi) + (1-\phi)\ln(1-\phi) + \chi\phi(1-\phi) + \left(w - \frac{1}{6}\right)\phi^3 + \frac{1}{24\pi g_\xi^3} - \frac{9}{16\pi}\frac{\left(\frac{1}{2} - \chi + w\phi\right)\phi}{\alpha^2 g_\xi}$$

Here, $\phi$ is the volume fraction (concentration) of the polymer in question, *r* is the chain length, $\chi$ (which is proportional to the second virial coefficient $B_2$), is the dimensionless Flory interaction parameter that quantifies the effective strengths of two-body interactions, *w* is a dimensionless coefficient that quantifies the effective strengths of three-body interactions and is proportional to the third virial coefficient, $g_\xi$ is proportional to $\xi$, and it quantifies the effective number of residues that contribute to correlation length, and the swelling ratio $\alpha$ quantifies the ratio of the chain dimensions in the dense droplet phase to the corresponding dimensions in the coexisting dispersed phase. We calculated binodals from Muthukumar's definition of the free energy of mixing, in conjunction with appropriate system-specific parameters, to construct binodals that best fit the experimental binodals that are associated with each of the five systems (LAF-1, LAF-1 + 3k poly-rA, LAF-1 + 30 poly-rA, LAF-1 + 15 poly-rA, and RGG). The fitting procedure involved using experimentally derived parameters (i.e., $\phi$ and $\chi$), setting $\alpha$ to 1.0 and calculating the $\chi$-dependent values of $g_\xi$. Our approach allows us to perform a single parameter fit to obtain the value of *w* for each of the five systems. Since the theory is laid out in terms of volume fractions, we first converted protein mass concentrations (mg/ml) into volume fractions ($\phi$) using a density conversion factor $\rho_0$.

$$\phi = c\left(\frac{v_m N_A}{M_m}\right) = \frac{c}{\rho_0}$$

Here, *c* is the mass concentration of protein, $M_m$ is the average mass of an amino acid (110 g/mol), $v_m$ is the average volume occupied by an amino acid (140 Å$^3$), $N_A$ is Avogadro's number, and $\rho_0$ is 1310.16 mg/mL. Using a single density factor is erroneous because it implies that the densities in the low and high concentration regimes are the same. This error is corrected by applying two separate constant offsets to the calculated low and high concentration arms of the binodals, which is also useful in that our analysis yields predictions regarding densities within the dispersed versus droplet phases for each construct.

We converted the measured $B_2$ values into $\chi$ using the following expression[4]

$$\chi = \frac{1}{2} - \left(\frac{M_2}{r}\right)^2 \left(\frac{1}{\underline{v_1}}\right) B_2$$

Here, $M_2$ is the mass of polymer (in kg/mol), *r* is the degree of polymerization (which we set to be the number of residues), and $\underline{v_1}$ is the partial molar volume of a monomer unit, which we define as 0.018 L/mol. Note that the *exact* value of the partial molar volume is inconsequential since we assume the same number for all constructs. This facilitates a direct conversion between the experimentally measured $B_2$ values and the values for $\chi$ at each salt concentration. We use these salt-dependent values of $\chi$ to construct binodals in a ($\phi$, $\chi$) space.

Having reconstructed the experimentally derived binodals in ($\phi$,$\chi$) space (solid symbols in Fig. 3b), we numerically fit the theoretical binodals from Muthukumar's theory that best match the experimental data for each system. This was achieved by using the free energy of mixing expression defined above to construct a free energy of mixing curve across all volume fractions for a given value of

$\chi$ and extracting the common tangent positions, which give rise to the low and high values of the coexisting volume fractions for the specific $\chi$. This process is repeated for the full range of $\chi$-values to generate the complete system-specific theoretical phase diagram. The process we use is similar to that described by Rubinstein and Colby[5] for the construction of binodals from the Flory-Huggins free energy of mixing expression. The fitting procedure outlined above was performed to search for the optimum value of the *w* parameter. To search for optimal, construct-specific, values of *w* we combined a swarm-based local non-linear optimization approach with random searching based on Monte Carlo sampling. We find that the functional form for *w vs.* goodness-of-fit parameter is convex with a globally optimal value obtained for all five constructs, thus giving us confidence in the estimated values of *w*.

Setting $\alpha$ to 1.0 assumes that the monomeric chain dimensions of LAF-1 and the RGG domain in the dilute phase are comparable to the chain dimensions in the dense phase. This is consistent with previous measurements on similar disordered proteins that undergo phase separation[6]. Importantly, however, we performed a parameter sensitivity analysis and found that the measured binodals are reproduced equally well when $\alpha$ lies in the interval of 0.7 to 1.2. This implies that it is highly unlikely that either LAF-1 or RGG undergo significant collapse ($\alpha$ considerably smaller than unity) or expansion ($\alpha$ considerably larger than unity) in droplets versus dispersed phases. Accordingly, it appears that small changes to chain dimensions are consistent with the measured phase behavior.

Three-body interactions refer to the impact of excluded volume due to the milieu or other chain molecules on the strengths of effective pairwise interactions between chains. If this coefficient is positive, then three-body interactions will have a diluting effect because they reduce the effective strengths of two body interactions. Accordingly, as *w* becomes more positive, the width of the two-phase regime narrows. Typically, the narrowing of the two-phase regime comes from more pronounced movement of the high concentration arm of the binodal. This is reasonable since the effects of three-body interactions should be more pronounced in environments of higher chain density. In contrast, changing the two-body interaction (as measured by $\chi$) has a more symmetrical effect on the low and high concentration arms of the binodal. Changes to either *w* or $\chi$ will alter the position of the critical point.

The parameter $g_\xi$ describes the correlation length in units of the number of residues. It is important to note that $g_\xi$ is *not* the same as the number of residues from a single continuous polymer-chain. It is instead the effective number of residues that contribute to a correlation length. This leads to a simple proportionality between $g_\xi$and $\xi$ where conversion of $g_\xi$ to $\xi$ is defined as follows:

$$\xi = \frac{g_\xi}{g_b} R$$

Here, $g_b$ defines the number of residues in a thermal blob – the length-scale over which an individual protein behaves as a Gaussian chain[7]. Previous work showed that a thermal blob spans ca. ~ 6-7 residues in sequences that are deficient in proline residues[8]. The parameter *R* refers to the size of a thermal blob. The radius of gyration, $\langle R_g \rangle$ and the end-to-end distance, $\langle R_{ee} \rangle$ of a thermal blob will bracket the value of *R*. The $\langle R_g \rangle$ of the thermal blob was previously determined to be 0.6 nm[8]. A thermal blob behaves as a Gaussian chain, and accordingly that the end-to-end distance can be written as:

$$\langle R_{ee} \rangle = \langle R_g \rangle \sqrt{6} = 0.6\sqrt{6} = 1.47 \text{ nm}$$

In Fig. 3c, we calculate $\xi$ by setting $R = \langle R_g \rangle$.

### 3. Justifications for co-opting the theory of binary mixtures

In fitting the experimentally measured binodals to Muthukumar's theory of polymer solutions, we assume that the system can be modeled as an effective binary mixture. For RGG and LAF-1 in the absence of RNA, this needs little justification, but when RNA is present we are effectively treating the RNA as a modulator of the two- and three-body interactions in solvent without taking it into account as an explicit macromolecule. This assumption is motivated by practical considerations (both experimental and theoretical) and because we obtain predictions for the parameters from the theory that have clear physical interpretations.

In order to obtain experimental data to explore the 3D titration of protein, RNA, and NaCl we would need to perform two-color coincidence experiment using the usFCS approach such that we label LAF-1 and the RNA to track the coincidence of these molecules. This would be essential to deconvolve the contributions of RNA-RNA and RNA-protein interactions and their salt dependencies vis-à-vis the salt dependence of the interactions between LAF-1 molecules. These types of experiments would be essential if the phase separation of LAF-1 required the presence of RNA. However, we find that LAF-1 drives phase separation without the requirement of RNA. Even the salt dependence of the low concentration arm of the LAF-1 binodal and the inferred critical point are not affected by the presence of RNA.

Second, standard Flory-Huggins theory or Flory-Huggins with a three-body correction term ($w$) cannot explain the phase behavior of LAF-1 and RGG in the absence of RNA. However, these data can be fit using Muthukumar's theory of polymer solutions, suggesting this might be the "simplest" two-component theory that can fit the data. Importantly, the data obtained in the presence of RNA are also well-fit using Muthukumar's theory. This fitting exercise indicates that the RNA molecules modulate the LAF-1 phase behavior mainly by increasing the magnitude of $w$ and altering the correlation length ($\xi$). These findings are consistent with the observed lack of significant changes to the salt dependence of the low concentration arm of the binodal of LAF-1 or the inferred critical point. Instead, it appears that the RNA molecules may crowd out the LAF-1 molecules within the droplet, thus shrinking the width of the two-phase regime for all salt concentrations.

Third, measurements of the mesh sizes *in vitro*, in the absence of RNA, agree with the predicted mesh size from theory. Upon the addition of RNA, our theoretical description suggests the mesh size should increase by ~2 nm. RNA is present *in vivo*, and based on dextran partitioning experiments the mesh size *in vivo* is ~2 nm larger than in the absence of RNA *in vitro*. Taken together, these results suggest that despite making a simplifying binary-mixture assumption, quantitative predictions from the resulting theory are confirmed when tested experimentally. Furthermore, for our *in vitro* experiments we used homopolymeric (poly-rA) monodisperse RNA at various lengths to interrogate the impact of RNA on the droplet's phase behavior. In reality, the collections of cellular RNAs span a diverse range of lengths, secondary and tertiary structure, and sequence motifs. The fact that our *in vitro* derived predictions hold true *in vivo* suggests that, despite the importance of RNA in modulating phase behavior, in the case of P-granules the mesh-size appears to be dictated by LAF-1. Conversely, our data do suggest that droplet dynamics are strongly impacted by longer RNA molecules.

Finally, Muthukumar's theory has not been generalized to ternary and higher order mixtures. This generalization is challenging since each of the polymer-specific parameters (three body interactions, swelling ratio, correlation length, etc.) will have a protein and RNA specific contribution. Importantly, even if such a theory were available, this would increase the number of experiments to be performed and a range of confounding difficulties would have to be addressed. Therefore, our use of the two-component theory represents a minimalist approach that yields good fits and physical insights. The ternary phase diagram would be a necessary future direction to explore in depth (see discussion section as well). Indeed understanding the impact of RNA sequence and structure on phase behavior represents a critical next step in relating phase behavior to biological function.

**4. Determining the second virial coefficient from the concentration dependence of diffusion**

For each construct, we estimated the second virial coefficient ($B_2$) by measuring the diffusion coefficients as a function of concentration in the dilute regime at several defined NaCl concentrations. For all salt concentrations, the measurements are made under conditions where the protein concentrations are always below the low concentration arm of the measured binodals, by up to an order of magnitude.

To determine $B_2$ from these data we use a formalism proposed by Harding and Johnson[9]. Plotting protein concentration (*c)* vs. diffusion constant (*D*) according to the following equation gives a straight line in the dilute regime;

$$D = D_0[1 + (2MB_2 - \bar{v} - k_s)c]$$

In addition to the parameters introduced above, *M* is the molar mass of the diffusing species, $\bar{v}$ is the partial molar volume of the solvent and $k_s$ is an empirical constant that accounts for the adjustments to the volume fraction that derive from the entrainment of the solvent along the polymer. When plotting *D* versus *c*, the slope of the line is equal to $(2B_2 - \bar{v} - k_s)$. For simplicity, we can define this term as;

$$k_D = 2MB_2 - \bar{v} - k_s$$

We make two key assumptions in using the equation of Harding and Johnson. First, we assume that the salt dependence of the observed diffusion coefficients derives mainly from the salt dependence of the solvent-mediated interactions between LAF-1 molecules. Accordingly, our definition of $k_D$ must be altered to;

$$k_D^{[\mathrm{NaCl}]} = 2MB_2^{[\mathrm{NaCl}]} - \bar{v} - k_s$$

As noted above, the partial molar volume of the solvent and the empirical constant $k_s$ (in units of mL /mg) quantify the degree of solvent entrainment. By assuming the form for $k_D$ that is shown in the equation above we are stipulating that $\bar{v}$ and $k_s$ do not strongly vary with increasing salt concentration. This assumption is justified by the recognition that all our FCS measurements are quite unlike the sedimentation velocity measurements in that they are made under dilute protein concentrations and away from the sedimentation regime, where the backflow of solvent upon sedimentation creates a problem with interpreting the impact on measured velocities in sedimentation velocity analytical ultracentrifugation experiments. Additionally, experiments show that the self-diffusion coefficient of

water decreases only by roughly 4% in the absence of NaCl when compared to the corresponding value in 1 M NaCl[10, 11]. This provides a proxy for estimating the extent of change to the solvent entrainment and suggests that the magnitude of the change we expect to the term will be at least two orders of magnitude smaller than the measured changes to the diffusion coefficient *D* as a function of protein concentration *c* (see Fig. 2b).

Our data show that the slope of the plot of *D vs. c* decreases as the salt concentration increases. We interpret this to imply a weakening of the homotypic associations between LAF-1/ RGG molecules. In the presence of high salt (1 M NaCl), we note that the diffusion coefficient *D* varies negligibly with protein concentration *c*. An example of the insensitivity of *D* to changes in protein concentration at high salt is shown in Fig. 2b. Based on these data we make the second assumption, which is that:

$$\bar{v} - k_s \approx 0$$

This assumption allows us to re-write the original expression as:

$$D = D_0 \left[1 + \left(2MB_{2,\mathrm{app}}^{[\mathrm{NaCl}]}\right) c\right]$$

Our assumptions are justified by our data and the magnitudes of the changes we observe to the measured diffusion coefficients as a function of protein concentration in different amounts of NaCl. In the equation above, our assumption involves the replacement of the actual second virial coefficient with a salt dependent apparent second virial coefficient, $B_{2,\mathrm{app}}^{[\mathrm{NaCl}]}$.

To assess the impact of such an assumption being incorrect, we re-examined the fitting procedure to assess how changes to $B_2$ might influence our results. While the absolute value associated with the inferred three-body coefficient *w* appears to change, the values of the inferred correlation lengths vary minimally with the absolute numerical value of $B_2$ (which is converted to $\chi$ in our analysis), but they instead depend mainly on the width of the measured two-phase regime (i.e., the actual measured binodals). Consequently, if there is a contribution from $\bar{v}$ and $k_s$, it does not change the derived correlation length.

The central determinant of our ability to fit our data to Muthukumar's theory originates from the shapes and numerical values for the concentrations along the binodal curves and the independent movements of the low and high concentration arms. These features are largely insensitive to the absolute value of $B_2$. Therefore, our assumptions are valid in the case of LAF-1, in part because we know *a priori* that it is strongly self-associative even at very low concentrations. For highly soluble proteins, where $B_2$ is unlikely to be the only factor in the concentration dependence of protein diffusion, these assumptions would not be expected to hold.

However, in order to test the veracity of the assumptions made to extract $B_2$ values from usFCS measurements, we compared the values obtained from right-angle laser light scattering. In this approach, one measures the concentration dependence of the scattered light of the solution with the protein as a function of protein concentration and subtracts the contribution from the buffer alone to uncover the second virial coefficient using a so-called Zimm plot (see Fig. S4). Here, the light source was a laser with wavelength, $\lambda$, of 488 nm with vertical polarization. Since the molecular size of each of the samples used was smaller than $\lambda/20$, no angular dependence for the excess scattered intensity was

expected and all light scattering data were recorded at an angle of 90°. The Rayleigh expression describing the intensity of light scattered from a particle in solution is given in

$$\frac{Kc}{R} = \frac{1}{M} + 2B_2c$$

where $K$ is an optical constant, $c$ is the particle concentration, $R$ is the Rayleigh ratio of scattered to incident light intensity, $M$ is the molecular weight, $B_2$ is the second virial coefficient. The optical constant is defined by

$$K = \frac{4\pi n^2 (dn/dc)^2}{N_A \lambda^4}$$

where $N_A$ is Avogadro's number, $n$ is the solvent refractive index, and $dn/dc$ is refractive index increment for the protein/solvent (~ 0.185 mL/g). The expression used to calculate the sample Rayleigh ratio, $R$, from a toluene standard is given as

$$R = \frac{I_A n_A^2 R_T}{I_T n_T^2}$$

where $I_A$ is the residual scattering intensity of the analyte (sample intensity – solvent intensity), $I_T$ is the toluene scattering intensity, $n$ is the solvent refractive index, $n_T$ is the toluene refractive index (1.503 at 488nm[12]), and $R_T$ is the Rayleigh ratio of toluene ($39.6 \times 10^{-6}$ $cm^{-1}$ at 488nm[13]).

We compared the measured $B_2$ values using right angle laser light scattering for LAF-1 at different salt concentrations to the values of $B_2$ that were obtained using usFCS measurements. The values were found to be equivalent across all salt concentrations. Therefore, for our analysis in the main text, we used data from usFCS measurements because these afford higher reliability at the low protein concentrations at which these measurements have to be made. Additionally, in contrast to light scattering, the impact of RNA molecules on $B_2$ can be readily quantified using usFCS measurements because the only signals in these measurements come from labeled molecules.

## 5. Concepts of overlap concentrations and correlation lengths

Polymer solutions can exist in one of four different concentration regimes: the dilute regime, the semidilute regime, the concentrated regime, and the polymer melt. In the dilute regime, individual polymer molecules exist as monomers in solution and experience minimal intermolecular interactions. In the semidilute regime, individual polymer molecules make fluctuating patterns of interactions with other chains. In the concentrated regime, polymers experience intermolecular entanglement, and the solution begins to experience viscoelastic bulk properties. Finally, in the polymer melt regime, the chains effectively solvate one another. A graphical representation of the dilute, semi-dilute and concentrated regimes is shown in Fig. S8.

The overlap volume fraction ($\phi^*$) or overlap concentration ($c^*$) defines the concentration at which inter-chain contacts become more likely than intra-chain contacts[5]. Practically, the overlap concentration defines the concentration threshold that separates the dilute and semi-dilute concentration regimes.

Formally, the overlap concentration is defined as

$$\phi^* = \frac{rv_m}{\frac{4}{3}\pi\langle R_{ee}\rangle^3}$$

Here $r$ is the degree of polymerization (i.e., the number of amino acids), $v_m$ is the average volume occupied by a single amino acid, and the denominator represents the volume of the pervaded volume calculated using the end-to-end distance, $R_{ee}$. The overlap volume fraction can be thought of as the instantaneous fraction of the pervaded volume occupied by the polymer at any given moment. The overlap concentration is defined as

$$c^* = \phi^* \rho_0$$

Here, $\rho_0$ is a density conversion factor that defines the mass concentration if the solution were 100% polymer; $\rho_0$=1310.16 mg/mL.

The overlap concentration can be determined experimentally, approximated using analytical models, or calculated using physical models that provide direct access to the end-to-end distance ($R_{ee}$). We used all-atom simulations to calculate the distribution of $R_{ee}$ values for LAF-1 RGG domain (see Fig. S7) and use this value to predict the overlap concentration. The LAF-1 RGG domain populates a heterogeneous ensemble of conformations characterized by large amplitudes of fluctuations. We generated an ensemble for 100,000 different conformations and for each conformation we calculated the overlap volume fraction. The resultant cumulative distribution function (CDF) allows us to quantify the probability of achieving a specific overlap concentration due to conformational fluctuations. These results are shown in Fig. S7b. The range of overlap volume fractions shown here is the same order of magnitude as the low concentration arms of the various binodals. The overlap volume fraction for the RGG domain is on the order of $10^{-3}$ to $10^{-2}$, which corresponds to an approximate overlap concentration of 1-10 mg/mL.

There is a well established relationship between the correlation length ($\xi$), the chain dimensions ($R_g$), and the polymer volume fraction ($\phi$). This was defined by de Gennes[14] as:

$$\xi \sim R_g \left(\frac{\phi}{\phi^*}\right)^x$$

Here $R_g$ is the radius of gyration of an individual polymer in dilute solutions, $\phi$ is the volume fraction of polymer, $\phi^*$ is the overlap concentration, and $x$ is a function of the polymer scaling exponent.

We can rewrite the above expression as:

$$\left(\frac{\phi}{\phi^*}\right)^x \sim \left(\frac{\xi}{R_g}\right)$$

For the RGG domain, based on the calculated overlap concentration and the measured low concentration binodal, it follows that:

$$\frac{\phi}{\phi^*} \approx 1;$$

Therefore;

$$\left(\frac{\phi}{\phi^*}\right)^x \approx 1$$

And finally,

$$\left(\frac{\xi}{R_g}\right) \approx 1;$$

As a result, when the low binodal and overlap concentration are approximately equal to one another then the correlation length should be on the same order of magnitude as each individual molecule.

## 6. Additional details regarding atomistic simulations

All-atom Monte Carlo simulations were performed using the CAMPARI software package (http://campari.sourceforge.net/) and the ABSINTH implicit solvent model[15]. In the ABSINTH model, protein atoms and solution ions are modeled explicitly, while the solvent is modeled using an implicit, mean-field representation with a dielectric constant set to 78.2. Conformation-specific atomic solvation states are calculated using solvent accessible volume fractions and references against experimentally determined solvation free energies for model compounds that make up polypeptides. Interaction potentials for van der Waals and electrostatic interactions amongst neutral groups are subject to 10 Å and 14 Å cutoffs, respectively. No cutoffs were imposed in calculations of electrostatic interactions between groups with a net charge as well as solution ions. Simulations were performed at 298 K in a spherical simulation droplet with a diameter of 43.2 nm to avoid the artifacts due to confinement.

Simulations were performed on the RGG domain of LAF-1 (residues 1-168) at two different salt concentrations at which phase separation would be expected to occur (neutralizing NaCl and 20 mM NaCl). Two hundred independent simulations were performed from unique starting conformations at each salt concentration, providing a total of $4 \times 10^9$ steps of simulation data. In addition to these main productions simulations, sets of simulations were run for 4 and 15 times longer than the main production simulations to assess overall convergence of statistics. In all cases, all parameters of interest showed nearly identical values irrespective of the total number of steps in any individual simulation run, suggesting that after an appropriate equilibration period the conformational ensemble generated represents a true equilibrium distribution. Simulation analysis was performed using analysis routines in CAMPARI, MDtraj[16], and CTraj, which is an in-house analysis package. Visualization was performed using VMD[17].

Simulations of atomistic self-avoiding random walk distributions of the RGG-domain, results of which are shown in Fig. S7a, were performed using CAMPARI, but setting the strength of all interaction terms to 0, with the exception of the repulsive component of the Lennard-Jones potential. This causes the self-avoiding RGG chain to have conformational statistics that are congruent with the excluded volume (EV) limit[18]. Simulations for the polyQ-160 system were performed in a similar manner to the LAF-1 RGG domain simulations.

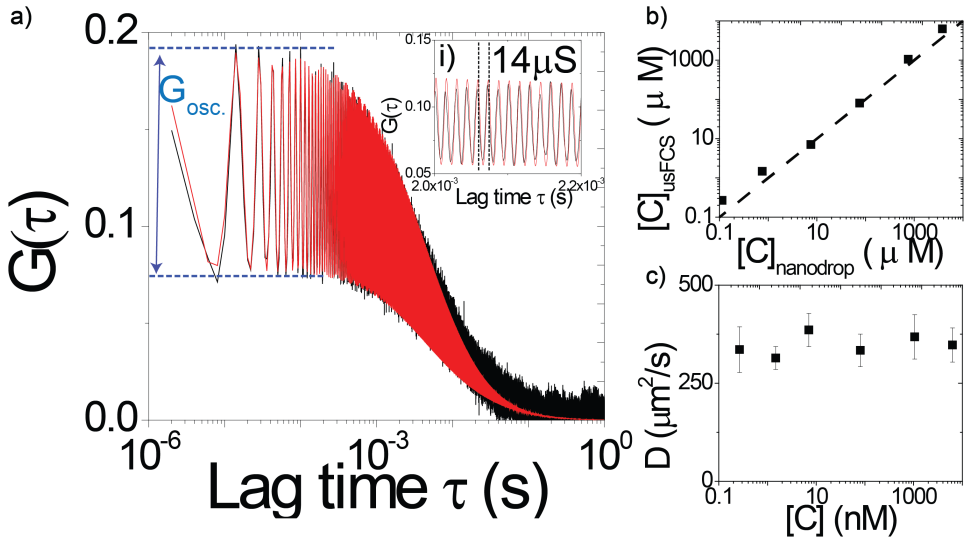


**Fig.S1** (a) Fluorescence autocorrelation of 14 nm hydrodynamics radius polystyrene particles while scanning at frequency 70 kHz. The fit to Equation 1 (red line) is shown. The magnitude of autocorrelation function oscillations with 14 µS period at short time scale, $G_{osc.}$, depends on the ratio between axially scanned distance ($Z$) and depth of focus ($\omega_z$). The inset shows fluorescence autocorrelation as a function of delay time ($\tau$) between 2 ms and 2.2 ms. The period (T) of autocorrelation curve is ~14 µS, which indicates the TAG lens scanning frequency ($T^{-1}$ ~70 kHz). When compared with standard FCS, ultra-fast-scanning FCS (usFCS) has several strengths. It increases the statistical accuracy for slowly moving molecules, by effectively sampling a larger volume. Improving this statistical accuracy allows for shorter measurement times than standard FCS, which helps to ensure accurate correlation curves. This approach also facilitates low excitation intensity, reducing the effect of photo-bleaching and optical saturation. Here, we show both (b) molecular concentration and (c) diffusivity from usFCS measurements as a function of Dylight 488 concentration.

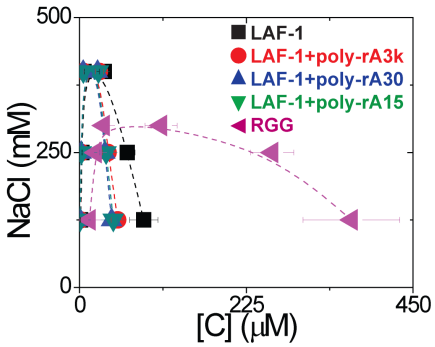


**Fig.S2** Phase diagram of LAF-1/NaCl concentrations, expressed in molar units of protein concentration. The left arms of the dashed coexistence curve indicate the phase boundary of dilute phase (protein concentrations outside droplets) and the right arms indicate the phase boundary of the condensed phase (protein concentrations inside droplets).

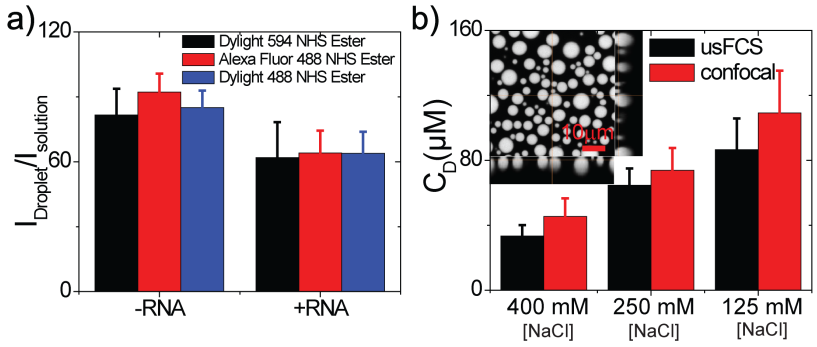


**Fig.S3** To assess the quantitative accuracy of our usFCS results, we used (a) three different fluorescent dyes to label LAF-1 and (b) determined the protein concentration using three-dimensional confocal microscopy. The consistent results obtained using different dyes imply that the fluorescent dyes do not have any effect on the phase behavior of LAF-1. We then measured the volume fraction ($\phi_{drop}$) of LAF-1 droplets using three-dimensional confocal microscopy to determine the protein concentration. To calculate the volume fraction, we first determined the both bulk ($C_B$) and saturation ($C_S$) protein concentrations using 280 UV absorption (Thermo nanodrop). Then, we calculated protein concentration in the droplet phase using $C_D = C_S + ((C_B - C_S)/ \phi_{drop})$ and compared the concentrations determined with the values obtained from usFCS, as shown in the right hand side panel above. The concentrations obtained using the methods are similar within error, confirming our results from the usFCS method.

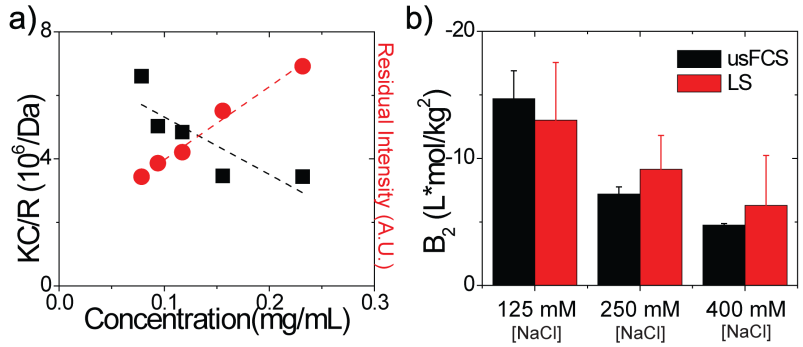


**Fig. S4** To test the quantitative accuracy of our estimates of second virial coefficients ($B_2$) obtained using usFCS based diffusion measurements that rely on two simplifying assumptions, we used right-angle laser light scattering to measure protein concentration dependence of the scattered light of the solution with the protein as a function of protein concentration to determine second virial coefficient. (a) Right-angle laser light scattering data for LAF-1 in 400 mM NaCl buffer solution. (b) Comparison of light scattering determined data with the values obtained from usFCS. The estimates of $B_2$ obtained using both methods are similar within error for three different salt concentrations, thus establishing the accuracy of our usFCS measurements and the validity of the assumptions used in our analysis of usFCS data.

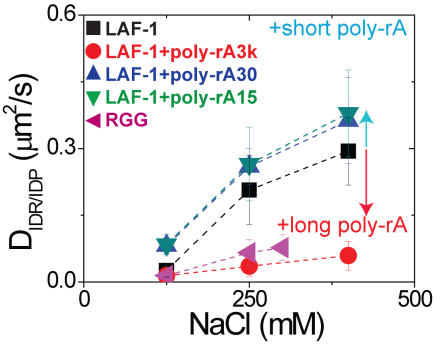


**Fig.S5** Increasing NaCl concentration increases the diffusion coefficients of LAF-1 in droplets. Adding short RNA (poly-rA30 and poly-rA15) also increases diffusion coefficient of LAF-1 in droplets whereas adding long RNA (poly-rA3k) decreases the diffusion coefficients of LAF-1 in droplets.

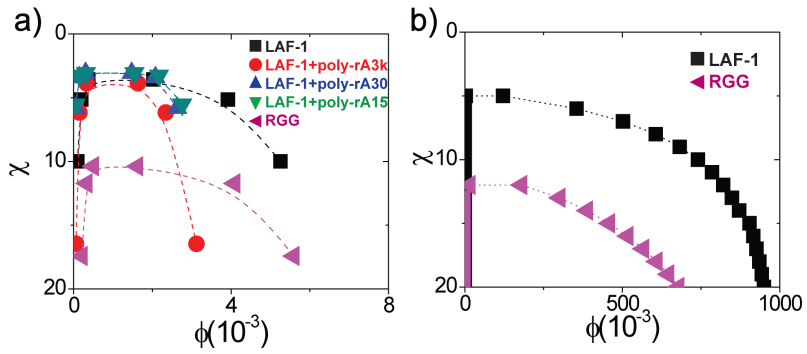


**Fig.S6** (a) The measured binodals are recast in terms of volume fractions. Here, we calculated Flory interaction parameters, $\chi$, converted from $B_2$ values and calculated protein volume fraction using average amino acid density ~1310.16 mg/mL. (b) The theoretical phase diagrams for LAF-1 and RGG that show the best reproduction obtainable with a simple Flory-Huggins style mean-field theory. Note that the resultant volume fractions for high concentration arms of binodals are almost three orders of magnitude greater than the measured counterparts.

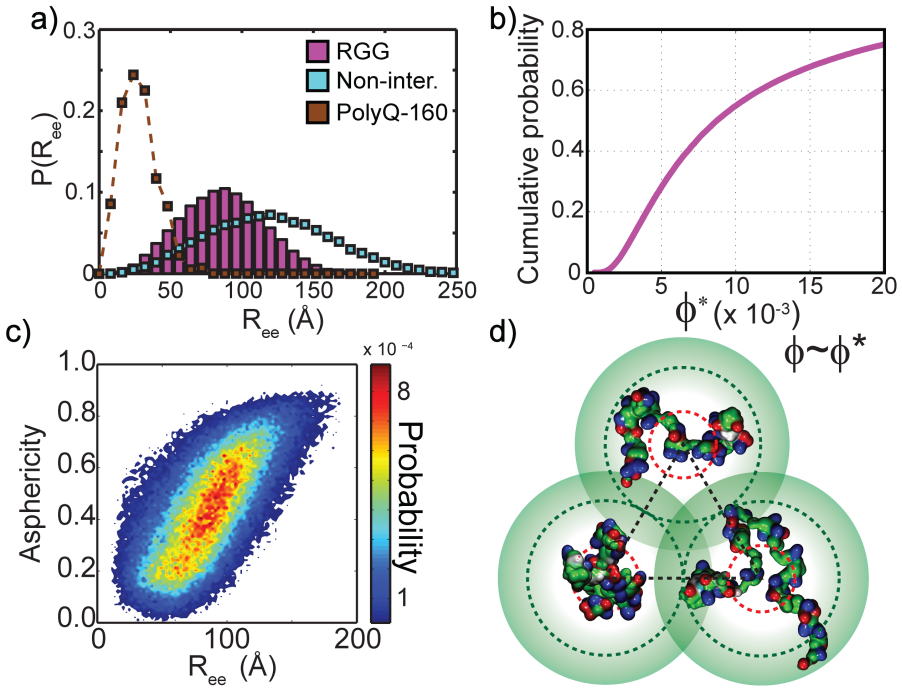


**Fig.S7** (a) The end-to-end distance, $R_{ee}$, distribution is shown for the RGG domain (RGG), a length-matched polyQ stretch (PolyQ-160), and an atomistic self-avoiding random walk version of the RGG domain that behaves as a non-overlapping random coil (Non-inter.). The polyQ chain forms dense, compact globules, giving rise to a substantially lower end-to-end distance distribution. The non-interacting chain is highly expanded due to the absence of any interactions that promote intra-chain interactions. The RGG ensemble samples conformations consistent with a highly collapsed globule as well as conformations consistent with a highly expanded chain. This chimeric conformational behavior is consistent with a chain that can form favorable inter- and intra-chain interactions while populating highly expanded ensembles. This result represents a biological manifestation of the 'stickers on a chain' heteropolymer that is congruent with the model of Semenov and Rubinstein[19]. (b) The overlap volume fraction ($\phi^*$), calculated for each conformation taken from the ensemble, allows us to compute the likelihood of crossing the overlap concentration. This is achieved by calculating the cumulative distribution function (CDF). The overlap volume fraction of the RGG domain, calculated from atomistic simulations, is in the same concentration regime as the experimentally measured and theoretically derived volume fractions for the low concentration arm. (c) The 2D histogram for the asphericity and end-to-end distance for the RGG ensemble were also calculated. The distribution shows a wide range of conformations explored, consistent with the conclusions drawn in panel (a). (d) Schematic illustration showing how the overlap concentration is impacted by the size of RGG domain. The calculated overlap concentration for the RGG domain is of the same magnitude as $c_D$, which corresponds to a number density of $2\text{x}10^{-4}$ molecules/nm$^3$ (i.e., 1 molecule every 17 nm shown in black dashed lines). The average size $R_{ee}$ of RGG is ~8.5 nm ($R_g$ ~ 3.5 nm shown in red dashed lines). RGG samples a broad range of conformations (~6 – 11 nm, envelope for which is shown as solid green contour). The RGG domain explores highly compact and highly expanded conformations, suggesting a low overlap volume fraction ($\phi^*$), which when combined with the large negative values of $B_2$, leads to a strong driving force for phase separation.

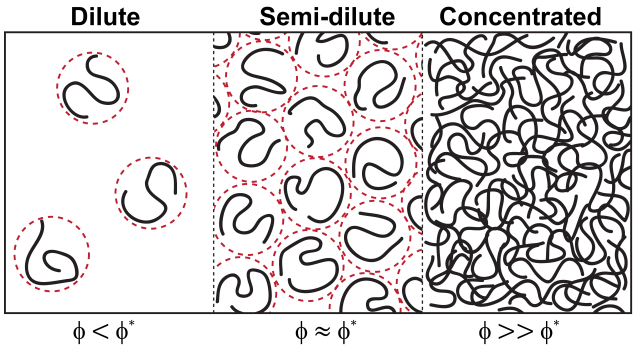


**Fig.S8** Graphical representation of the dilute, semi-dilute and concentrated regimes. Here, $\phi$ represents the polymer concentration in the solution and $\phi^*$ represents the overlap volume fraction. LAF-1 droplets are consistent with a semi-dilute solution.

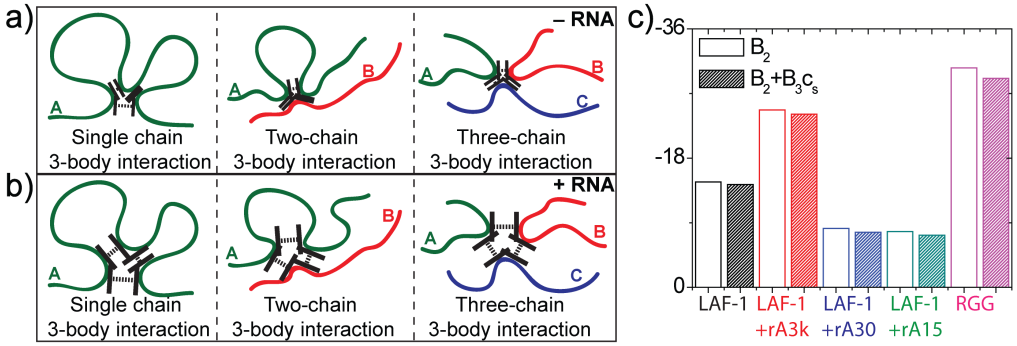


**Fig. S9** Graphical depiction of three categories of three-body interactions with (a) and without (b) RNA. The "elbows" show contacts between a pair of points, the dashed line in each elbow depicts a repulsive interaction, and the length of the dashed line quantifies the effective strength of a contact. Importantly, a three-body interaction represents the cumulative effect of the three elbows. Three-way contacts can be made within a single chain (left most sub-panel); they could involve a pair of chains (central sub-panel) where two of the contact points come from a single chain; or three chains could come together to make the three-body interactions (right most sub-panel on the top row). Not shown here are the two-body interactions, which for all cases in this work are attractive. The three-body interaction dilutes the effect of the attractive two-body interaction, but does not out-compete it i.e., the net inter-chain interaction is attractive. Upon the addition of RNA the three-body interaction becomes stronger leading to a further dilution of the two-body interaction strength. This manifests itself as a reduction in the concentration of the high concentration arm of the binodal. (c) Quantification of the effects of the three-body term on the effective strengths of two-body interactions at the equilibrium protein concentration outside the droplet, $c_s$. The strengths of three-body interactions, $w$, were converted into third virial coefficients, $B_3$ using the following expression: $B_3 = w/2(r/M)^3 v_c^2$. Here, $M$ is the mass of protein, $r$ is the degree of polymerization (i.e., the number of residues), and $v_c$ is the partial molar volume of a monomer unit (defined as 0.018 l/mol). Comparison of $B_2$ with $B_2$+$B_3c_S$ shows that the strengths of three-body interactions with positive values of $w$ lead to a weakening pairwise interactions.

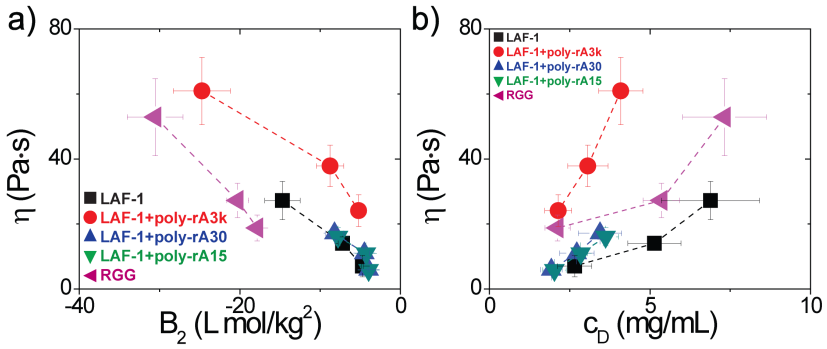


**Fig. S10** Viscosity of condensed droplets as a function of (a) intra-molecular interaction, $B_2$ and (b) protein concentration within droplets, $c_D$.

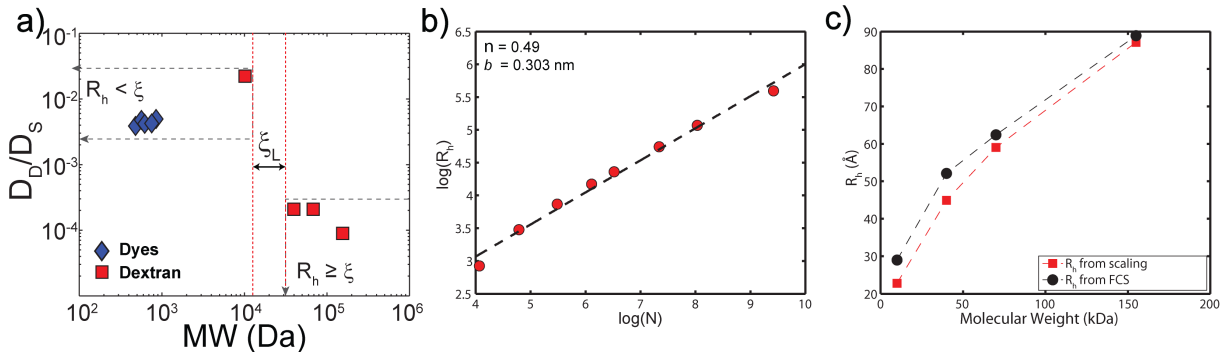


**Fig. S11** Using polymeric probes in the form of dextran molecules of differing molecular weights to determine the effective mesh size. (a) Unlike spherical probes, dextran molecules are flexible polymers. For molecular weights below a threshold value, dextran molecules should behave like small solutes[20]. Conversely, above a threshold molecular weight, denoted as $M_{w,\xi}$, the diffusivities of dextran molecules should decrease as the reciprocal of increasing molecular weight[20]. We measured the diffusivities of dextran molecules of four different molecular weights in droplets ($D_D$) and in bulk solution ($D_S$). Following the theoretical analysis of Cai *et al.*[20], we plotted the ratios of $D_D$ to $D_S$ as a function of molecular weight. This analysis shows that the ratio of diffusion coefficients for the 10 kDa dextran molecule is within an order of magnitude of the values for small fluorescent dyes even though the molecular weights are different by two orders of magnitude. In contrast, beyond a threshold molecular weight of 40 kDa, the ratio of diffusivities for dextran molecules decrease by at least two orders of magnitude and show a decrease with increased molecular weight dependence that is expected of polymeric probes. Based on these results, the dextran molecular weight that corresponds to the lower bound of the mesh size lies between 10 kDa and 40 kDa. The results are in line with theoretical predictions, and point to the existence of a threshold lower bound of the mesh-size, delineated by $\xi_L$. (b) Published data in the literature show that dextran behaves as an ideal chain in aqueous buffers[21]. To verify this, we analyzed published data [22] for hydrodynamic radii as a function of molecular weight. The measurements were based on viscosity analysis and light scattering. Here, we plot data from the literature as a log-log plot, with the log of the hydrodynamic radius ($R_h$) along the ordinate and the log of

the degree of polymerization ($N$) along the abscissa. This analysis yields a straight line with a slope that corresponds to the scaling exponent $v$ and an intercept that corresponds to log($b$), where $b$ is the Kuhn length. We find $v$ = 0.49, which is similar, within finite size considerations, to the theoretical exponent of 0.5 expected for an ideal chain. The intercept yields a value of $b$ = 0.303 nm. For molecular weights below a threshold value, dextran molecules should behave like small solutes[20]. Conversely, above a threshold molecular weight, denoted as $M_{w,\xi}$, the diffusivities of Dextran molecules should decrease as the reciprocal of increasing molecular weight [20]. We measured the diffusivities of dextran molecules of four different molecular weights in droplets ($D_D$) and in bulk solution ($D_S$). Following the theoretical analysis of Cai *et al.*[20] for non-sticky probes, we plotted the ratios of $D_D$ to $D_S$ as a function of molecular weight. This analysis shows that the ratio of diffusion coefficients for the 10 kDa Dextran molecule is within an order of magnitude of the values for small fluorescent dyes even though the molecular weights are different by two orders of magnitude. In contrast, beyond a threshold molecular weight of 40 kDa, the ratio of diffusivities for dextran molecules decrease by at least two orders of magnitude and show a decrease with increased molecular weight dependence that is expected of polymeric probes. Setting, $M_{w,\xi}$ to be 40 kDa yields a value of $N_\xi \approx 245$ for the degree of polymerization of dextran probes that match the effective degree of polymerization of the mesh corresponding to the intra-droplet polymer network. Using $N_\xi$ we estimate $\xi$ to be $b(N_\xi)^{0.5} \approx 4.7$ nm. Here, we use $b$ = 0.303 nm, which is the measured Kuhn length of dextran molecules. To probe the robustness of these estimates, we performed partitioning experiments using the dextran molecules. (c) In the main text, we used $R_h$ values to plot apparent viscosities as a function of probes size. To put this analysis on a quantitative footing, we used the scaling analysis extracted from published data to back-calculate the hydrodynamic radii predicted by this scaling behavior for the dextran molecular weights used in our study, and compared those values to the hydrodynamic radii determined by FCS. These two methods yielded results that agree with one another.

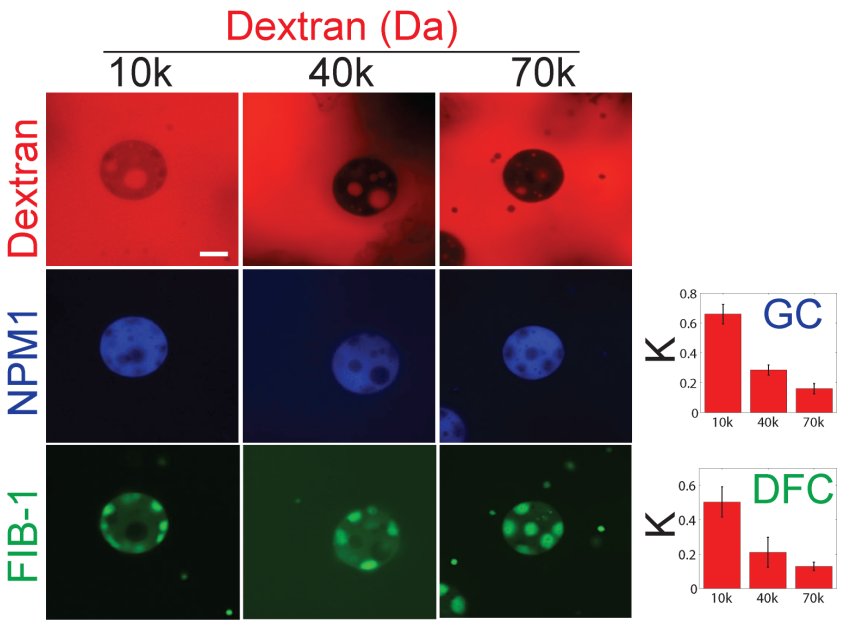

**Fig. S12** Dextran (1 mg/mL, red) was directly injected into nuclei of *X. laevis* oocytes expressing NPM1::CFP (blue) and FIB1::GFP (green), scale bar = 10 µm. Nuclei were dissected in mineral oil 1-2 hours after microinjection and imaged using laser scanning confocal microscopy under *in vivo* conditions. Partition coefficients (K) were calculated from background corrected fluorescent intensities inside/outside either the granular component (GC) marked by NPM1::CFP or the dense fibrillar component (DFC) marked by FIB1::GFP.

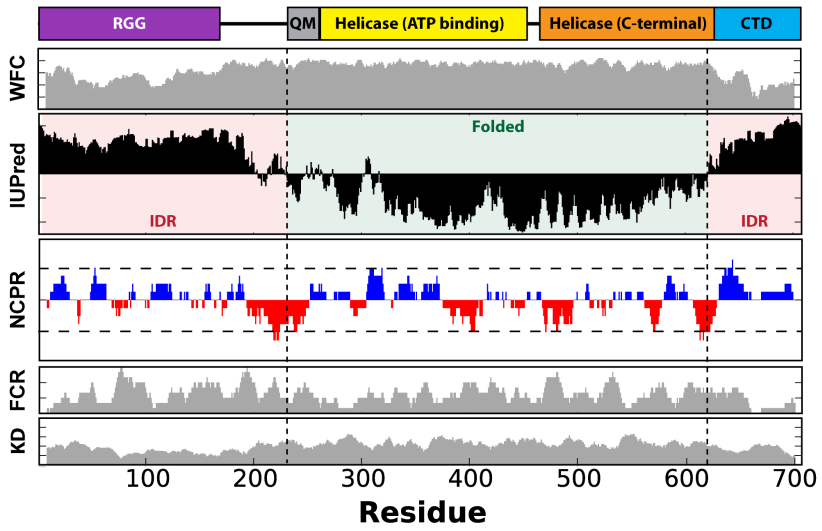


**Fig.S13** Linear sequence analysis of LAF-1. Each track describes a difference type of sequence feature to provide a general summary of the linear amino acid sequence. **WFC** represents Wootton-Federhen sequence complexity[23]; the RGG and CTD domains show a significantly reduced complexity when compared to the remainder of the sequence. **IUPred** represents the predicted disorder score based on the IUPred algorithm[24]; N-terminal and C-terminal intrinsically disordered regions (IDRs) are identified. **NCPR** represents the linear net charge per residue[25]; the N-terminal domain shows both net positively charged local regions and net negatively charged regions, suggesting electrostatic interactions may play a role in driving RGG-RGG interaction. **FCR** represents the fraction of charged residues; several regions in the RGG contain a high FCR despite minimal net charge, indicating these regions have the characteristics of a strong polyampholyte. **KD** represents the Kyte-Doolite hydrophobicity scale[26]; the folded domains are significantly more hydrophobic than the IDRs.

**References**


1. Holehouse AS, Das RK, Ahad JN, Richardson MO, Pappu RV. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal* 2017, **112**(1)**:** 16-21.
2. Muthukumar M. Thermodynamics of polymer solutions. *Journal of chemical physics* 1986, **85**(8)**:** 4722-4728.
3. Flory PI. Thermodynamics of high polymer solutions. *Journal of Chemical Physics* 1942, **10**(1)**:** 51-61.
4. Cowie MG, Arrighi V. *Polymers: Chemistry and Physics of Modern Materials*. CRC Press: Boca Raton, 2007.

5. Rubinstein M, Colby RH. *Polymer Physics*. Oxford University PRess: New York, 2003.
6. Burke KA, Janke AM, Rhine CL, Fawzi NL. Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Molecular Cell* 2015, **60**(2)**:** 231-241.
7. de Gennes PG. *Scaling concepts in polymer physics*. Cornell University Press: Ithaca, N.Y., 1979.
8. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(33)**:** 13392-13397.
9. Harding SE, Johnson P. The concentration-dependence of macromolecular parameters. *Biochemical Journal* 1985, **231**(3)**:** 543-547.
10. Ben Ishai P, Mamontov E, Nickels JD, Sokolov AP. Influence of Ions on Water Diffusion 瑆 A Neutron Scattering Study. *The Journal of Physical Chemistry B* 2013, **117**(25)**:** 7724-7728.
11. Müller K, Hertz H. A parameter as an indicator for water-water association in solutions of strong electrolytes. *The Journal of Physical Chemistry* 1996, **100**(4)**:** 1256-1265.
12. Moutzouris K, Papamichael M, Betsis SC, Stavrakas I, Hloupis G, Triantis D. Refractive, dispersive and thermo-optic properties of twelve organic solvents in the visible and near-infrared. *Applied Physics B-Lasers and Optics* 2014, **116**(3)**:** 617-622.
13. Bender TM, Lewis RJ, Pecora R. Absolute Rayleigh ratios of four solvents at 488 nm. *Macromolecules* 1986, **19**(1)**:** 244–245.
14. Degennes PG. Dynamics of entangled polymer solutions. I. The Rouse model. *Macromolecules* 1976, **9**(4)**:** 587-593.
15. Vitalis A, Pappu RV. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *Journal of Computational Chemistry* 2009, **30**(5)**:** 673-699.
16. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX*, et al.* MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* 2015, **109**(8)**:** 1528-1532.
17. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* 1996, **14**(1)**:** 33-38.
18. Mao AH, Lyle N, Pappu RV. Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochemical Journal* 2013, **449:** 307-318.
19. Semenov AN, Rubinstein M. Thermoreversible gelation in solutions of associative polymers. 1. Statics. *Macromolecules* 1998, **31**(4)**:** 1373-1385.
20. Cai L-H, Panyukov S, Rubinstein M. Mobility of Nonsticky Nanoparticles in Polymer Liquids. *Macromolecules* 2011, **44**(19)**:** 7853-7863.
21. Antoniou E, Tsianou M. Solution properties of dextran in water and in formamide. *Journal of Applied Polymer Science* 2012, **125**(3)**:** 1681-1692.
22. Armstrong J, Wenby R, Meiselman H, Fisher T. The hydrodynamic radii of macromolecules and their effect on red blood cell aggregation. *Biophysical journal* 2004, **87**(6)**:** 4259-4270.
23. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry* 1993, **17**(2)**:** 149-163.
24. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005, **21**(16)**:** 3433-3434.
25. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences* 2010, **107**(18)**:** 8183-8188.
26. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* 1982, **157**(1)**:** 105-132.